Amendments to the specification

Please replace the paragraph beginning on page 9, line 5 with the following amended version of that paragraph:

An introduction to genetic algorithms can be found in David E. Goldberg (1989) Genetic Algorithms in Search, Optimization and Machine Learning Addison-Wesley Pub Co; ISBN: 0201157675 and in Timothy Masters (1993) Practical Neural Network Recipes in C++ (Book&Disk edition) Academic Pr; ISBN: 0124790402. A variety of more recent references discuss the use of genetic algorithms used to solve a variety of difficult problems. *See*, e.g., ~~http://~~garage.cse.msu.edu/papers/papers-index.html and the references cited therein; ~~http://~~gaslab.cs.unr.edu/ and the references cited therein; ~~http://~~www.aic.nrl.navy.mil/ and the references cited therein; ~~http://~~www.cs.gmu.edu/research/gag/ and the references cited therein and ~~http://~~www.cs.gmu.edu/research/gag/pubs.html and the references cited therein.

Please replace the paragraph beginning on page 9, line 30 with the following amended version of that paragraph:

Sequences of those polynucleotides found to have desired characteristics are deconvoluted (e.g., sequenced, or, when positional information is available, by noting the position of the polynucleotide). This is performed by DNA sequencing, by reading a position on an array, real time PCR (e.g., ~~TaqMan~~TAQMAN), restriction enzyme digestion, or any other method noted herein, or currently available.

Please replace the paragraph beginning on page 10, line 19 with the following amended version of that paragraph:

Example Advantages of GAGGS

There are a variety of advantages to GAGGS as compared to the prior art. For example, physical access to genes/organisms is not required for GAGGS, as sequence information is used for oligo design and selection. A variety of public databases provide extensive sequence information, including, e.g., ~~Genbank~~GENBANK™ and those noted *supra*. Additional sequence databases are available on a contract basis from a variety of companies specializing in genomic information generation and storage.

Please replace the paragraph beginning on page 16, line 26 with the following amended version of that paragraph:

One example algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul *et al., J. Mol. Biol.* 215:403-410 (1990). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (~~http://~~www.ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al., supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff (1989) *Proc. Natl. Acad. Sci. USA* 89:10915).

Please replace the paragraph beginning on page 20, line 8 with the following amended version of that paragraph:

For example, oligonucleotides *e.g.,* for use in *in vitro* amplification/ gene reconstruction methods, for use as gene probes, or as shuffling targets (e.g., synthetic genes or gene segments) are typically synthesized chemically according to the solid phase phosphoramidite triester method described by Beaucage and Caruthers (1981), *Tetrahedron Letts.*, 22(20):1859-1862, *e.g.,* using an automated synthesizer, as described in Needham-VanDevanter *et al.* (1984) *Nucleic Acids Res.*, 12:6159-6168. Oligonucleotides can also be custom made and ordered from a variety of commercial sources known to persons of skill. There are many commercial providers of oligo synthesis services, and thus this is a broadly accessible technology. Any nucleic acid can be custom ordered from any of a variety of commercial sources, such as The Midland Certified Reagent Company (mcrc@oligos.com), The Great American Gene Company (~~http://~~www.genco.com), ExpressGen Inc. (www.expressgen.com), Operon Technologies Inc. (Alameda, CA) and many others. Similarly, peptides and antibodies can be custom ordered from any of a variety of sources, such as PeptidoGenic (pkim@ccnet.com), HTI Bio-products, inc. (~~http://~~www.htibio.com), BMA Biomedicals Ltd (U.K.), Bio·Synthesis, Inc., and many others.

Please replace the paragraph beginning on page 36, line 11 with the following amended version of that paragraph:

Essentially any nucleic acid can be shuffled using the GAGGS methods herein. No attempt is made herein to identify the hundreds of thousands of known nucleic acids. Common sequence repositories for known proteins include ~~GenBank~~ GENBANK EMBL, DDBJ and the NCBI. Other repositories can easily be identified by searching the internet.

Please replace the paragraph beginning on page 43, line 24 with the following amended version of that paragraph:

For example, neural net approaches can be coupled to genetic algorithm-type programming. for example, NNUGA (Neural Network Using Genetic Algorithms) is an available program (~~http://~~www.cs.bgu.ac.il/~omri/NNUGA/) which couples neural networks and genetic algorithms. An introduction to neural networks can be found, e.g., in Kevin Gurney (1999) An Introduction to Neural Networks, UCL Press, 1 Gunpowder Square, London EC4A 3DE, UK. and at ~~http://~~www.shef.ac.uk/psychology/gurney/notes/index.html. Additional useful neural network references include those noted above in regard to genetic algorithms and, e.g., Christopher M. Bishop (1995) Neural Networks for Pattern Recognition Oxford Univ Press; ISBN: 0198538642; Brian D. Ripley, N. L. Hjort (Contributor) (1995) Pattern Recognition and Neural Networks Cambridge Univ Pr (Short); ISBN: 0521460867.

Please replace the paragraph beginning on page 44, line 2 with the following amended version of that paragraph:

A protein design cycle, involving cycling between theory and experiment, has led to recent advances in rational protein design. A reductionist approach, in which protein positions are classified by their local environments, has aided development of appropriate energy expressions. Protein design programs can be used to build or modify proteins with any selected set of design criteria. *See*, e.g., ~~http://~~www.mayo.caltech.edu/; Gordon and Mayo (1999) "Branch-and-Terminate: A Combinatorial Optimization Algorithm for Protein Design" Structure with Folding and Design 7(9):1089-1098; Street and Mayo (1999) "Intrinsic ß-sheet Propensities Result from van der Waals Interactions Between Side Chains and the Local Backbone" Proc. Natl. Acad. Sci. USA, 96, 9074-9076; Gordon et al. (1999) "Energy Functions for Protein Design" Current Opinion in Structural Biology 9(4):509-513 Street and Mayo (1999) "Computational Protein Design" Structure with Folding and Design 7(5):R105-R109; Strop and Mayo (1999) "Rubredoxin Variant Folds Without Iron" J. Am. Chem. Soc. 121(11):2341-2345; Gordon and Mayo (1998) "Radical Performance Enhancements for Combinatorial Optimization Algorithms based on the Dead-End Elimination Theorem" J. Comp. Chem 19:1505-1514; Malakauskas and Mayo (1998) "Design, Structure, and Stability of a Hyperthermophilic Protein Variant" Nature Struct. Biol. 5:470. Street and Mayo (1998) "Pairwise Calculation of Protein Solvent-Accessible Surface Areas" Folding & Design 3: 253-258. Dahiyat and Mayo (1997) "De Novo Protein Design: Fully Automated Sequence Selection" Science 278:82-87; Dahiyat and Mayo (1997) "Probing the Role of Packing Specificity in Protein Design" Proc. Natl. Acad. Sci. USA 94:10172-10177; Dahiyat et al. (1997) "Automated Design of the Surface Positions of Protein Helices" Prot. Sci. 6:1333-1337; Dahiyat et al. (1997) "De Novo Protein Design: Towards Fully Automated Sequence Selection" J. Mol. Biol. 273:789-796; and Haney et al. (1997) "Structural basis for thermostability and identification of potential active site residues for adenylate kinases from the archaeal genus *Methanococcus*" Proteins 28(1):117-30. These design methods rely generally on energy expressions to evaluate the quality of different amino acid sequences for target protein structures. In any case, designed or modified proteins or character strings corresponding to proteins can be directly shuffled in silico, or reverse translated and shuffled in silico and/or by physical shuffling. Thus, one aspect of the invention is the coupling of high-throughput rational design and in silico or physical shuffling and screening of genes to produce activities of interest.

Please replace the paragraph beginning on page 45, line 1 with the following amended version of that paragraph:

Similarly, molecular dynamic simulations such as those above and, e.g., Ornstein et al. (~~http://~~www.emsl.pnl.gov:2080/homes/tms/bms.html; Curr Opin Struct Biol (1999) 9(4):509-13) provide for "rational" enzyme redesign by biomolecular modeling & simulation to find new enzymatic forms that would otherwise have a low probability of evolving biologically.

For example, rational redesign of p450 cytochromes and alkane dehalogenase enzymes are a target of current rational design efforts. Any rationally designed protein (e.g., new p450 homologues or new alkaline dehydrogenase proteins) can be evolved by reverse translation and shuffling against either other designed proteins or against related natural homologous enzymes. Details on p450s can be found in Ortiz de Montellano (ed.) 1995, Cytochrome P450 Structure and Mechanism and Biochemistry, Second Edition Plenum Press (New York and London).

Please replace the paragraph beginning on page 49, line 1 with the following amended version of that paragraph:

An overall issue for the above strategy is the availability of enough related sequences generated through shuffling to provide useful information. An alternative to shuffling sequences is to apply the modeling tools to all available sequences, e.g., the ~~GenBank~~ GENBANK database and other public sources. Although this entails massive computational power, current technologies make the approach feasible. Mapping all available sequences provides an indication of sequence space regions of interest. In addition, the information can be used as a filter which is applied to in silico shuffling events to determine which virtual progeny are preferred candidates for physical implementation (e.g., synthesis and/or recombination as noted herein).

Please replace the paragraph beginning on page 51, line 18 with the following amended version of that paragraph:

HMM can be used in other ways as well. Instead of applying the generated profile to identify previously unidentified family members, the HMM profile can be used as a template to generate de novo family members (e.g., intermediate members of a cladistic tree of nucleic acids). For example, the program, HMMER is available (~~http://~~hmmer.wustl.edu/). This program builds a HMM profile on a defined set of family members. A sub-program, HMMEMIT, reads the profile and constructs de novo sequences based on that. The original purpose of HMMEMIT is to generate positive controls for the search pattern, but the program can be adapted to the present invention by using the output as in silico generated progeny of a HMM profile defined shuffling. According to the present invention, oligonucleotides corresponding to these nucleic acids are generated for recombination, gene reconstruction and screening.

Please replace the paragraph beginning on page 54, line 28 with the following amended version of that paragraph:

Alternatively, real time quantitative PCR (e.g., ~~TaqMan~~ TAQMAN) can be performed where PCR oligos are highly discriminating for the feature of interest. This can be done by, for example, having a polymorphism unique to the motif present at or near the 3' end of an oligo such that it will only prime the PCR efficiently if there is a perfect match. Real time PCR product analysis by, e.g., FRET or ~~TaqMan~~ TAQMAN (and related real time reverse-transcription PCR) is a family of known techniques for real time PCR monitoring that has been used in a variety of contexts (*see*, Laurendeau et al. (1999) "TaqMan PCR-based gene dosage assay for predictive testing in individuals from a cancer family with INK4 locus haploinsufficiency" Clin Chem 45(7):982-6; Laurendeau et al. (1999) "Quantitation of MYC gene expression in sporadic breast tumors with a real-time reverse transcription-PCR assay" Clin Chem 59(12):2759-65; and Kreuzer et al. (1999) "LightCycler technology for the quantitation of bcr/abl fusion transcripts" Cancer Research 59(13):3171-4.

Please replace the paragraph beginning on page 59, line 8 with the following amended version of that paragraph:

Typically, PDA starts with a protein backbone structure and designs the amino acid sequence to modify the protein's properties, while maintaining it's three dimensional folding properties. Large numbers of sequences can be manipulated using PDA, allowing for the design of protein structures (sequences, subsequences, etc.). PDA is described in a number of publications, including, e.g., Malakauskas and Mayo (1998) "Design, Structure and Stability of a Hyperthermophilic Protein Variant" Nature Struc. Biol. 5:470; Dahiyat and Mayo (1997) "De Novo Protein Design: Fully Automated Sequence Selection" Science, 278, 82-87. DeGrado, (1997) "Proteins from Scratch" Science, 278:80-81; Dahiyat, Sarisky and Mayo (1997) "De Novo Protein Design: Towards Fully Automated Sequence Selection" J. Mol. Biol. 273:789-796; Dahiyat and Mayo (1997) "Probing the Role of Packing Specificity in Protein Design" Proc. Natl. Acad. Sci. USA, 94:10172-10177; Hellinga (1997) "Rational Protein Design – Combining Theory and Experiment" Proc. Natl. Acad. Sci. USA, 94:10015-10017; Su and Mayo (1997)" Coupling Backbone Flexibility and Amino Acid Sequence Selection in Protein Design" Prot. Sci. 6:1701-1707; Dahiyat, Gordon and Mayo (1997) "Automated Design of the Surface Positions of Protein Helices" Prot. Sci., 6:1333-1337; Dahiyat and Mayo (1996) "Protein Design Automation" Prot. Sci., 5:895-903. Additional details regarding PDA are available, e.g., at ~~http://~~www.xencor.com/.

Please replace the paragraph beginning on page 67, line 4 with the following amended version of that paragraph:

Similarly, PRINTS (e.g., Atwood et al., *above*) is a compendium of protein motif fingerprints derived from the OWL composite sequence database. Fingerprints are groups of motifs within sequence alignments whose conserved nature allows them to be used as signatures of family membership. Fingerprints can provide improved diagnostic reliability over single motif methods by virtue of the mutual context provided by motif neighbors. The database is now accessible via the UCL Bioinformatics Server on ~~http:@~~ www.biochem.ucl.ac.uk/bsm/dbbrowser/. Atwood et al. describe the database, its compilation and interrogation software, and its Web interface. *See also*, Attwood et al. (1997) "Novel developments with the PRINTS protein fingerprint database" Nucleic Acids Res 25(1):212-7.

Please replace the paragraph beginning on page 74, line 5 with the following amended version of that paragraph:

One approach to screening diverse libraries is to use a massively parallel solid-phase procedure to screen cells expressing shuffled nucleic acids, e.g., which encode enzymes for enhanced activity. Massively parallel solid-phase screening apparatus using absorption, fluorescence, or FRET are available. *See*, e.g., United States Patent 5,914,245 to Bylina, et al. (1999); *see also*, ~~http://~~www.kairos-scientific.com/; Youvan et al. (1999) "Fluorescence Imaging Micro-Spectrophotometer (FIMS)" Biotechnology et alia[[<]] www.et-al.com[[>]] 1:1-16; Yang et al. (1998) "High Resolution Imaging Microscope (HIRIM)" Biotechnology et alia, [[<]]www.et-al.com[[>]] 4:1-20; and Youvan et al. (1999) "Calibration of Fluorescence Resonance Energy Transfer in Microscopy Using Genetically Engineered GFP Derivatives on Nickel Chelating Beads" posted at www.kairos-scientific.com. Following screening by these techniques, sequences of interest are typically isolated, optionally sequenced and the sequences used as set forth herein to design new sequences for in silico or other shuffling methods.

Please replace the paragraph beginning on page 75, line 5 with the following amended version of that paragraph:

A variety of commercially available peripheral equipment and software is available for digitizing, storing and analyzing a digitized video or digitized optical or other assay images, *e.g.*, using PC (Intel x86 or ~~pentium~~ PENTIUM chip- compatible DOS™, OS2™ WINDOWS™, WINDOWS NT™ or WINDOWS95™ based machines), MACINTOSH™, or UNIX based (*e.g.*, SUN™ work station) computers.

Please replace the paragraph beginning on page 75, line 21 with the following amended version of that paragraph:

Current art computational hardware resources are fully adequate for practical use in GAGGS (any mid-range priced ~~Unix~~UNIX system (e.g., for ~~Sun Microsystems~~SUN MICROSYSTEMS) or even higher end ~~Macintosh~~MACINTOSH or PCs will suffice). Current art in software technology is adequate (i.e., there are a multitude of mature programming languages and source code suppliers) for design of an upgradable open-architecture object-oriented genetic algorithm package, specialized for GAGGS users with a biological background.

Please replace the paragraph beginning on page 76, line 1 with the following amended version of that paragraph:

For example, standard desktop applications such as word processing software (e.g., ~~Microsoft Word~~MICROSOFT WORD™ or ~~Corel WordPerfect~~COREL WORDPERFECT™) and database software (e.g., spreadsheet software such as ~~Microsoft Excel~~MICROSOFT EXCEL™, ~~Corel Quattro Pro~~COREL QUATTRO PRO™, or database programs such as ~~Microsoft Access~~MICROSOFT ACCESS™ or ~~Paradox~~PARADOX™) can be adapted to the present invention by inputting one or more character string into the software which is loaded into the memory of a digital system, and performing a GO as noted herein on the character string. For example, systems can include the foregoing software having the appropriate character string information, e.g., used in conjunction with a user interface (e.g., a GUI in a standard operating system such as a ~~Windows~~WINDOWS, ~~Macintosh~~MACINTOSH or LINUX system) to manipulate strings of characters, with GOs being programmed into the applications, or with the GOs being performed manually by the user (or both). As noted, specialized alignment programs such as PILEUP and BLAST can also be incorporated into the systems of the invention, e.g., for alignment of nucleic acids or proteins (or corresponding character strings) as a preparatory step to performing an additional GO on the resulting aligned sequences. Software for performing PCA can also be included in the digital system.

Please replace the paragraph beginning on page 78, line 26 with the following amended version of that paragraph:

In one internet embodiment, a client system typically executes a Web browser and is coupled to a server computer executing a Web server. The Web browser is typically a program such as IBM's ~~Web Explorer~~ WEB EXPLORER, ~~Internet explorer~~INTERNET EXPLORER, ~~NetScape~~NETSCAPE or ~~Mosaic~~MOSAIC. The Web server is typically, but not necessarily, a program such as IBM's HTTP Daemon or other WWW daemon (e.g., LINUX-based forms of the program). The client computer is bi-directionally coupled with the server computer over a line or via a wireless system. In turn, the server computer is bi-directionally coupled with a website (server hosting the website) providing access to software implementing the methods of this invention.

Please replace the paragraph beginning on page 80, line 10 with the following amended version of that paragraph:

The functions to encode two or more biological molecules can provide one or more windows wherein the user can insert representation(s) of biological molecules. In addition, the encoding function also, optionally, provides access to private and/or public databases accessible through a local network and/or the intranet whereby one or more sequences contained in the databases can be input into the methods of this invention. Thus, for example, in one embodiment, where the end user inputs a nucleic acid sequenced into the encoding function, the user can, optionally, have the ability to request a search of ~~GenBank~~ GENBANK and input one or more of the sequences returned by such a search into the encoding and/or diversity generating function.

Please replace the paragraph beginning on page 81, line 15 with the following amended version of that paragraph:

Generally the charts are schematics of arrangements for components, and of process decision tree structures. It is apparent that many modifications of this particular arrangement for DEGAGGS, e.g., as set forth herein, can be developed and practiced. Certain quality control modules and links, as well as most of the generic artificial neural network learning components are omitted for clarity, but will be apparent to one of skill. The charts are in a continuous arrangement, each connectable head-to tail. Additional material and implementation of individual GO modules, and many arrangements of GOs in working sequences and trees, as used in GAGGS, are available in various software packages. Suitable references describing exemplar existing software are found, e.g., at ~~http://~~www.aic.nrl.navy.mil/galist/ and at ~~http://~~www.cs.purdue.edu/ coast/archive/clife/FAQ/www/Q20_2.htm. It will be apparent that many of the decision steps represented in Figs. 1-4 are performed most easily with the assistance of a computer, using one or more software program to facilitate selection/ decision processes.